Integrative analysis identifies immune-related enhancers and lncRNAs perturbed by genetic variants associated with Alzheimers disease

Alexandre Amlie-Wolf[1,2,3], Mitchell Tang[2,3], Jessica King[2,3], Beth Dombroski[2,3], Yi-Fan Chou[2,3], Elizabeth Mlynarski[2,3], Gerard D. Schellenberg[1,2,3], Li-San Wang[1,2,3]

1) Genomics and Computational Biology Graduate Group, Perelman School of Medicine; 2) Institute for Biomedical Informatics, Perelman School of Medicine; 3) Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, Perelman School of Medicine; **University of Pennsylvania**, Philadelphia, PA, USA
alexaml@upenn.edu – http://tesla.pcbi.upenn.edu/∼alexaml/

Introduction

Dozens of genetic variants associated with late-onset Alzheimer's disease (LOAD) have been identified by genome-wide association studies (GWAS). However, these are only tag markers for nearby genetic variants in linkage disequilibrium (LD) and may not be actually functional. Moreover, all 21 of the significant variants identified in phase 1 of the International Genomics of Alzheimer's Project (IGAP) meta-analysis [1] are in non-protein-coding regions, implicating gene regulatory mechanisms as underlying the association signals. These considerations suggest a need for functional annotation of expanded sets of variants spanning the LD blocks tagged by the IGAP variants in order to identify the truly causal variants, their effects on regulatory mechanisms, the tissue context of this regulation, the affected target genes, and the direction of these effects on gene expression.

To address this need, we developed a novel tool, called INFERNO (INFERring the molecular mechanisms of NOncoding genetic variants). Given a list of tagging variants, INFERNO uses 1,000 Genomes Project [2] data to define LD blocks. Expanded sets of variants are annotated with:

- Sites of enhancer RNA (eRNA) transcription across 112 tissue facets from FANTOM5 [3]
- ChromHMM-defined epigenetic enhancer states across 127 tissues and cell types from Roadmap Epigenomics [4]
- Transcription factor binding sites (TFBSs) for 332 transcription factors predicted by HOMER [5]

Tissues and cell types from each data source are grouped into 32 broad tissue categories for cross-data source comparison, and empirical p-values for the enrichment of functional overlaps in each tissue category and tag region are obtained by background sampling. Co-localization analysis is then performed to identify shared causal signals underlying both the IGAP GWAS signals and GTEx expression quantitative trait loci across 44 tissues [6] in order to characterize the affected target genes and tissue contexts.

Application of INFERNO to Alzheimer's Disease Genetic Signals

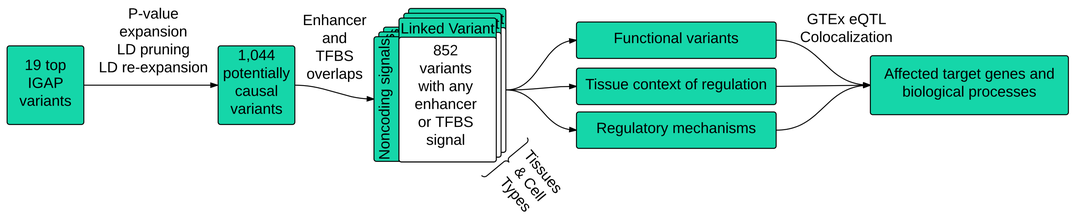

Figure 1: Flowchart of analysis approach

We applied INFERNO to 19 noncoding variants associated with late-onset Alzheimer's Disease (LOAD) identified in phase 1 of IGAP, excluding the variant in the DSG2 region, which did not replicate, and the variant in the HLA region, which is notoriously hard to analyze. We defined an expanded set of 496 variants by identifying all variants within 500 kb of any tag SNP with p-value within one order of magnitude of the top p-value. Then we subjected this set to LD pruning, yielding 52 variants, which were submitted as input to the INFERNO tool. After LD re-expansion, 1,044 unique variants were analyzed for regulatory potential.

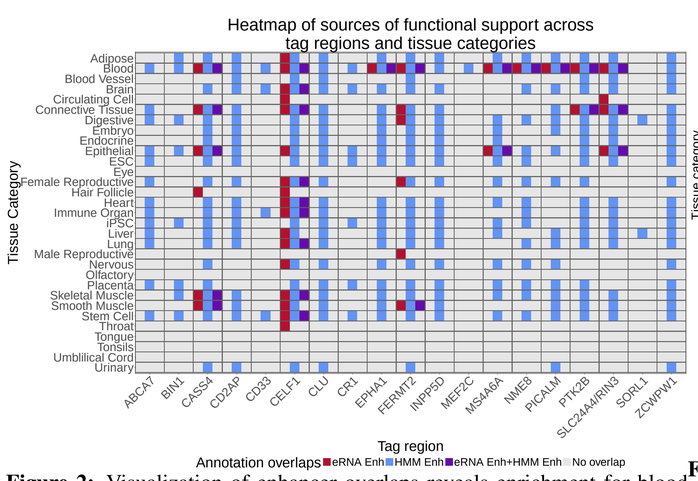

Figure 2: Visualization of enhancer overlaps reveals enrichment for blood category overlaps, implicating immune activity. eRNA Enh: FANTOM5 Enhancer. HMM Enh: Roadmap Enhancer state defined by ChromHMM

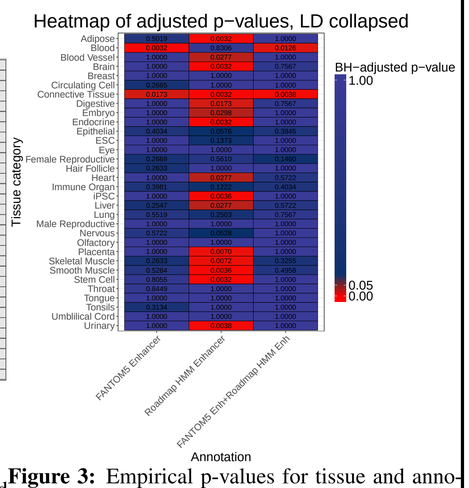

Figure 3: Empirical p-values for tissue and annotation overlap combinations based on 10,000 background samples supports blood and connective tissue (fibroblast) signals

Co-localization analysis with GTEx eQTLs

Direct overlap with GTEx eQTL data found 750 variants across 16 tag regions that were significant eQTLs, but these direct overlaps are subject to LD biases, so we instead used the COLOC Bayesian method [7] to identify GWAS and eQTL signals sharing a causal variant (H_4). We applied COLOC to eQTL signals for 876 unique genes across all 19 tag regions (median number of genes tested in each region = 33) for a total of 24,963 tests of GWAS - eQTL co-localization. This identified 154 sets of tag regions, tissues, and target genes with a high probability ($P(H_4) >= 0.5$) of having a shared causal signal, across 15 tag regions, 37 tissues, and 67 target genes. This model also provides Approximate Bayes Factors (ABFs) representing the probability that a given variant is the shared causal variant, which we use in addition to TFBS overlap to prioritize individual variants for validation.

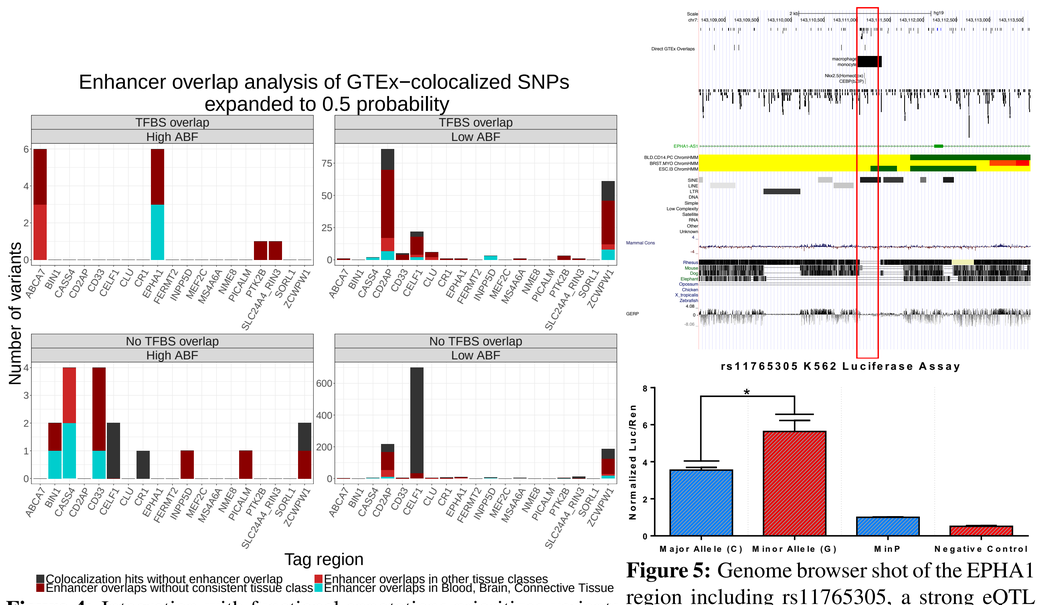

Figure 4: Integration with functional annotations prioritizes variants overlapping enhancers from concordant tissue classes with motif overlaps, high ABF values, or both

Figure 5: Genome browser shot of the EPHA1 region including rs11765305, a strong eQTL for the EPHA1-AS1 lncRNA, and luciferase assay results in K562 cells, a leukemia of monocyte precursor cell line. Asterisk represents statistically significant luciferase expression at the 0.05 level

Summary table of top prioritized results

Tag Region	Affected mechanism and evidence	Direction of effect
ABCA7	Digestive system regulation of ABCA7 and CNN2, high ABF variant	Risk allele, ↑ABCA7 and CNN2 expression
BIN1	Lymphocyte regulation of BIN1, high ABF variant	Risk allele, ↑ BIN1 expression
CASS4	HOXD13-mediated enhancer with blood eQTL for CASS4, High ABF variant for digestive CASS4 eQTL	Protective allele, ↓ blood, ↑ fibroblasts
CD2AP	Strong homeobox TF disruption in enhancer for RP11-385F7.1 in blood and brain, affecting GTPase signaling	Risk allele, ↓ brain, other tissues
CD33	Whole blood regulation of CD33, tag variant colocalized with eQTL, high ABF variant	Protective allele ↓ CD33 expression
CELF1	Brain signal for RP11-750H9.5, moderate TF disruption, affecting immune regulatory hub	Risk allele, ↓ lncRNAs
EPHA1	Very strong ABF for rs11765305 affecting EPHA1-AS1 (→ JAK2) and two taste receptor signals in blood (monocytes) with strengthened CEBP motif	Protective allele, ↑↑ EPHA1-AS1 expression
INPP5D	Blood signal for INPP5D, strong disruption of Homeobox TFs and moderate on other TFs	Risk allele, ↓ INPP5D expression
ZCWPW1	One SNP strongly disrupts several motifs and colocalizes with GTEx brain eQTLs for GAL3ST4, PVRIG, and STAG3	Protective allele, varying regulatory effects

Conclusions

The INFERNO tool provides an easy and powerful approach for inferring the molecular mechanisms of noncoding genetic variants. We have implemented INFERNO in an efficient pipeline with source code and access to a web server version that will be available at http://lisanwanglab.org/INFERNO.

The application of INFERNO to the analysis of LOAD-associated noncoding genetic signals identified a small number of putatively causal variants with strong functional evidence, and the significant enrichment of functional overlaps in the blood and connective tissue categories supports the hypothesis of immune activity as an important aspect of LOAD pathology.

Methods

INFERNO is implemented using Python, R, and bash. Datasets from each consortium were grouped into tissue categories based on the categorization provided by Roadmap and the CL ontology. For bootstrapping, variants were matched on minor allele frequency (bin size 0.01), distance to the nearest TSS (rounded to 1kb), and the number of LD partners. Multiple testing correction was performed using the Benjamini-Hochberg procedure. Co-localization analysis used the COLOC R package and a custom script to analyze genes tested with each tag variant across all GTEx tissues.

References

[1] J C Lambert, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–8, 2013.

[2] Adam Auton, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, sep 2015.

[3] Robin Andersson, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61, mar 2014.

[4] Roadmap Epigenomics Consortium, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, 2015.

[5] Sven Heinz, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, 2010.

[6] K. G. Ardlie, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, may 2015.

[7] Claudia Giambartolomei, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5), 2014.

Acknowledgements

This work is supported by NIA T32-AG00255, NIGMS R01-GM099962, NIA U24-AG041689, U01-AG032984, and P30-AG10124. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. We gratefully acknowledge Casey Brown, Barbara Engelhardt, Mingyao Li, Eddie Lee, Fanny Leung, Pavel Kuksa, Barry Slaff, and Nikolaos Vrettos for feedback and support.